Deloitte.

Together makes progress

Building a modern data center strategy in the AI era: **Turning volatility into advantage**

Introduction

For years, the data center was treated as an overhead cost by organizations: A necessary, but not strategic, component of the organization. That framing no longer fits. The rapid rise of artificial intelligence (AI) has meant an organization's data center strategy and footprint has become either a competitive advantage or disadvantage for the organization. Today, the data center determines how quickly you can ship, how reliably you can run, and how confidently you can invest.

Most organizations have a hybrid data center posture, but one that is hybrid-by-default; a patchwork of cloud contracts, colocation sites, and on-premise facilities, with upgrades assembled under pressure. It is the intentionality upon which this hybrid landscape is built that matters, and our underlying thesis is that companies, so often hybrid-by-default, should consider a more intentional, hybrid-by-design strategy for their data center footprint, due to the competitive advantages that such strategy can bring to the organizations. Companies should think of the estate as a managed portfolio: public cloud, on-premises sites, edge locations, and colocation for high-density AI.

This article discusses how the move to hybrid-by-design replaces ad hoc choices with explicit placement rules, unit-level economics, and a visible exceptions process. And even where hybrid-by-design is already in place, there is usually more to do to make the data center an asset: standardize patterns across pillars, simplify tooling and handoffs, surface costs at the service level, and retire legacy on schedule.

AI is defining the requirements of this data center managed portfolio. Models demand more capacity and tighter proximity to data. Power and cooling matter in new ways as the sheer physical requirements demanded by AI exponentially increase. Leaders are required to make rapid placement calls: what stays close to customers, what sits next to data, what moves to specialized compute, and how those choices show up in speed, risk, and unit cost.

That is why foundations come first: resiliency, standardization, and simplification. Resiliency means cyber, application, and infrastructure protections that work as one, so recovery is predictable everywhere. Standardization means repeatable patterns and controls, so teams do not waste cycles reinventing them. Simplification cuts tools and handoffs that slow delivery and increase risk.

From there, the data center strategy shifts to infrastructure-first. Orthodoxy has been to start from an application-based approach. We propose starting from the infrastructure portfolio—the full mix of cloud, on-premises, edge, and colocation capacity the business depends on. Ask which outcomes are we optimizing? Time to market? Control and compliance? Recovery targets? Cost per unit of work? After defining outcomes, we map capacity to where it delivers best. A transparent decision tree, reviewed on a fixed cadence, ensures shifts in demand and AI growth translate rapidly into action.

Setting up your modern data center strategy

Guiding principles for hybrid-by-design

Three principles guide the shift from hybrid-by-default to the intentional, hybrid-by-design strategy we suggest.

First, co-ownership with the business. Create a standing governance forum that includes the chief information officer (CIO) or chief technology officer (CTO), chief financial officer (CFO), chief operations officer (COO), risk and security leaders, and major product heads. Grant this group decision rights over workload placement, source-of-service choices, and decommission milestones. Frame every choice in board-relevant terms such as margin improvement, speed to market, earnings impact, and risk reduction. Make AI-driven placement and capacity decisions explicit in this forum so model demand, power/cooling headroom, and data governance are addressed up front.

Second, resilience and value must be considered together. Design placement, economics, and compliance as simultaneous, not competing, criteria. A balanced scorecard should track both resilience metrics, such as availability, recovery, sovereignty conformance, and incident frequency, and value metrics, such as time to AI use case, margin lift, cost per transaction, and cycle time to usable data set.

Third, ensure service economics are transparent. Shift funding models from infrastructure line items to outcomes business owners can see and control. Publish unit costs such as cost per transaction or per model call. For AI services, publish per-model and per-token costs and the service-level objectives (SLOs) that govern them, so owners can make informed placement trade-offs. Apply showback or chargeback and reinvestment rules that tie savings directly to modernization.

The decision framework: Where and how to run workloads

Begin with a common framework so that every decision ties directly to outcomes the business will recognize. Apply these **five lenses** to create a shared view and define the target state.

1. **AI-enabled** and traditional enterprise workloads. Start from the work itself. Classify AI patterns (training, fine-tuning, inference, embedded) and traditional enterprise workloads (systems of record, batch, analytics). For each class, set targets for latency to users and data, control and compliance, recovery objectives, and unit cost. Use those targets to drive placement, capacity, and data adjacency across the portfolio. Connect each class to a clear business outcome such as time to AI use case, customer experience, regulatory posture, or margin so decisions stay anchored to results rather than preferences.

2. **Application portfolio.** Build a current inventory and map applications to value, consumption, and risk. Use a 6R lens to retire low-value assets and consolidate duplicates. Move bursty services to autoscaling or event-driven patterns. Free budget and graphics processing unit (GPU) capacity for priority AI work. Produce a modernization backlog with owners, expected benefits, and dates so changes land in a managed sequence.

3. **Data layer and governance.** Classify sensitive data at creation and maintain that context as workloads migrate. Define residency and sovereignty rules that placement decisions must meet. Standardize monitoring, policy checks, and evidence capture across platforms so that compliance requirements are met without slowing delivery. Treat model artifacts and training data sets as regulated assets.

4. **Infrastructure footprint.** Inventory assets across on-premises, colocation, cloud, and edge environments. Surface performance, latency, and capacity hot spots, and address quick wins such as rightsizing and autoscaling. Document power and cooling headroom, network proximity to data sources, and cross-connect limits for AI-intensive zones, so sourcing decisions reflect both cost and operating realities.

5. **IT operating model.**
   Re-benchmark spans of control, organizational layers, and team topology. Align platform and product teams to business capabilities rather than infrastructure silos. Clarify decision rights and SLO ownership for top workloads so that work moves without unnecessary handoffs. Ensure SLO ownership explicitly covers AI services (training, inference) and accelerator capacity planning. The outcome is an operating model that delivers speed, compliance, and clear accountability.

From these five lenses, organizations should produce a set of target principles, a placement policy, a modernization backlog with expected benefits, a cost and risk baseline, and a balanced scorecard that pairs resilience and value measures for executive review.

Architecture principles: Resilience and value by design

Translate the framework into principles that guide every placement and modernization decision. The intent is resilience and value at the same time.

Placement policy comes first.
Make workload placement decisions using a concise set of criteria: latency, control requirements, residency and sovereignty obligations, unit-level economics, and recovery objectives. Reject one-size-fits-all approaches. Any exceptions should be visible, formally approved, and documented with risk acceptance. Map training, fine-tuning, inference, and embedded AI to distinct placement profiles that reflect latency, control, and recovery needs.

Infrastructure as a business decision

A global financial institution operating aging on-premises facilities with limited visibility into cost, risk, and performance faced rising AI demand and tighter regulatory obligations. To adopt a hybrid-by-design model, the enterprise focused on baselining TCO and SLOs; mapping data residency and access requirements; establishing a concise placement policy with a clear exceptions process; instrumenting service-level telemetry and unit costs; sequencing modernization with decommission milestones; and provisioning high-density AI capacity in colocation while keeping sensitive systems in controlled sites. Quarterly portfolio reviews tied placement choices to business metrics: time to market, recovery targets, and cost per transaction/model call.

In practice, they started by creating one clear view of cost, risk, and performance, then set simple rules for where work should run and why. Changes rolled out in waves: stand up AI-ready capacity where it made sense, keep sensitive systems where control was essential, and shut down legacy capacity on schedule. With a quarterly review and a plain-English price list for platform services, leaders could see trade-offs, measure progress, and redirect savings to the next set of improvements.

Segment where it creates advantage.
Establish AI-intensive zones where proximity to data, accelerator density, power, cooling, and network fabric are critical. Maintain general business services in estates designed for predictable availability and cost. This separation simplifies capacity planning and reduces contention as AI demand accelerates.

Create transparent service economics.
Make costs observable at the service and workload level so teams can respond. Apply showback or chargeback models, rightsizing, autoscaling, and serverless for spiky demand. Publish unit costs (e.g., cost per transaction or per model call) so product owners make placement choices based on value delivered, not simply on the lowest expense line.

Build governance into the fabric.
Apply consistent controls across cloud, colocation, on-premises, and edge environments. Classify data at creation, preserve tags and policies as workloads move, and automate monitoring and evidence capture. Incorporate vendor resilience and third-party risk into design so that failover and recovery are not only tested but also properly funded.

AI has become the stress test for every infrastructure decision and is exposing where policies, economics, or capacity lag behind ambition.

Making it real: Economics, execution, and talent

Economics (investment and TCO governance)

Traditional total-cost-of-ownership (TCO) frameworks miss the reality of AI. Unlike prior technology waves governed by licenses or virtual machines, **AI spend—measured in tokens—often scales in nonlinear and unpredictable ways**, and technical decisions can drive token cost and implications for the modern data center. In this way, cost should be treated as a design input. Build one view of AI costs that the business understands in unit terms, so placement and funding choices can be made before code moves.

1. **Build one comparable TCO model.** Cover cloud, colocation, and on-prem, including facilities, power and cooling, depreciation, licenses, network, labor, security, compliance, and cost of capital. Express results as cost per transaction, per model call, per gigabyte (GB) processed, or per user served.

2. **Decide funding with scenarios** (see sidebar, "Four future scenarios"). Stress test the model against an acquisition that doubles volume, an AI pilot that becomes a production service, or new residency rules. Compare total and unit costs for retrofit, build and own, and partner options before committing.

3. **Expose and retire transition costs.** Make dual running visible, time bound, and funded. Tie each wave to decommission milestones and confirm that contracts, licenses, cross-connects, and backups are retired when new capacity goes live.

4. **Publish a clear rate card.** Provide a simple price list for platform services mapped to unit metrics. Review with product owners and the CFO, and codify how verified savings are reinvested in the modernization backlog.

Four future scenarios

The next decade of data center strategy will be shaped by two defining uncertainties: the pace of AI adoption and the economics of compute.

In the **Edge of Everything** future, high AI adoption combines with decentralized compute. AI becomes ambient and compute becomes borderless. Tokenized, decentralized networks such as Bittensor and Render gain traction as enterprises seek flexible, low-cost capacity. Micro data centers and edge clusters proliferate. The implication is to design for modularity, local inference, and integration with open compute protocols.

In the **GigaCore Era**, high AI adoption combines with hyperscaler dominance. AI becomes a utility sold by the gigaflop as hyperscalers build vast AI superclusters and colocate near major power sources. The implication is to secure long-term access to power, standardize on hyperscaler chip formats, and align with public-private AI infrastructure initiatives.

In **Open Gridlock**, low AI adoption combines with decentralized compute. Decentralized capacity outpaces demand. Protocols mature technically but fail to reach scale. The implication is to stay capital light, monetize excess capacity, and experiment with infrastructure-as-a-service offerings for emerging tokenized models.

In **Steady Giants**, low AI adoption combines with hyperscaler dominance. AI remains a productivity tool rather than a transformation. Growth stabilizes and hyperscalers focus on efficiency and compliance. The implication is to prioritize sustainability leadership, optimize legacy facilities, and manage return-on-investment expectations for AI-focused builds.

Across all four future scenarios, a hybrid-by-design strategy turns volatility into advantage, enabling organizations to thrive whether the future tilts toward hyperscale concentration or decentralized flexibility.

5. **Set economic guardrails.** Require tagging, rightsize capacity, enable autoscaling where appropriate, schedule nonproduction environments, use serverless or event-driven patterns for bursty demand, and enforce storage tiering and life cycle policies. Improve TCO without undermining SLOs.

6. **Avoid common traps.** Do not chase cheaper regions if latency or residency policies will be violated. Avoid long commitments that outrun demand forecasts. In colocation, model cross-connect and power pricing over the term, not just space. In cloud, include egress and managed service dependencies. Keep any exception to placement policy visible with documented risk acceptance.

Execution: Roadmap and scorecard

Organizations should treat modernization as a managed program with clear sequencing, decision rights, and checkpoints the business co-owns. The intent is to capture quick savings, remove structural constraints, and create an operating rhythm that adapts to AI demand, regulation, and market shifts.

Near term (0–2 quarters)

- Baseline TCO and SLOs
- Turn off idle capacity
- Enforce tagging and a simple rate card
- Stand up cross-platform governance with named owners
- Publish placement policy and exceptions process
- Define the first AI-ready zone and harden monitoring

Mid term (2–6 quarters)

- Modernize priority apps using a 6R lens
- Apply autoscaling or event-driven patterns for bursty demand
- Rebalance placement to cut dual running
- Set partner guardrails and validate site security and continuity
- Start decommission waves and retire contracts and licenses

Long term (6 or more quarters)

- Segment footprints for AI-intensive and business operations
- Build a sustainable capacity plan with partners
- Exit legacy on schedule
- Embed multi-site recovery and vendor resilience
- Continue reskilling and selective hiring for platform, site reliability engineering (SRE), and AI operations

Using a single scorecard allows leaders to see resilience and value on the same page. Keep measures few, comparable across platforms, and tied to decisions the business understands. Review monthly for telemetry and quarterly with product owners and finance, then refresh targets annually as AI ambition and regulation change.

Resilience metrics	Value metrics
• Critical service availability and error rate	• Time to production for new AI use cases
• Mean time to detect and mean time to recover	• Revenue or margin lift from top data-enabled product
• Recovery objectives met by service, including tested failover	• Cost per transaction or per model call, with trend
• Policy conformance for data residency and access control	• Cycle time from data request to usable data set
• Third-party risk posture and continuity test results	• Percent idle capacity eliminated and reinvested
• Security incident rate and time to contain	• Customer experience signals; for example, latency and abandonment

How to use it

Assign an executive owner for each metric. Set target and tolerance bands, define the corrective action when a band is breached, and link improvements to funding and decommission milestones. Show both current level and trajectory so leaders can judge if changes are durable. Keep exceptions visible with documented risk acceptance, and require that partner service-level agreements (SLAs) align to your targets so recovery and performance remain consistent across the footprint.

Talent

Architecture will not deliver results without the right people and ways of working. Organizations should re-benchmark how teams are structured and align them to business capabilities rather than infrastructure silos. Platform engineering, SRE, and AI operations should be established as durable functions, while product teams should be made accountable for SLOs on their most critical workloads. Decision rights should be clarified so that placement, security, and cost choices move quickly through clear ownership and small review loops.

Workforce transition must be planned deliberately. Many legacy data center teams are later in their careers and take pride in what they have built, which makes reskilling a genuine change program. Organizations should provide defined pathways into SRE, platform engineering, and AI operations through structured training, shadowing opportunities, and incentives that reward new skills. When external hiring is required, leaders should anticipate paying a premium for cloud and AI expertise and compete directly with technology firms. Partners can be used selectively to fill peaks in demand while internal capability matures.

Security and compliance should be integrated into daily work rather than applied as an afterthought. Controls and monitoring should operate consistently across cloud, colocation, on-premises, and edge environments so that teams are not forced to reinvent policy for each setting. Performance management should be tied to the outcomes boards care about most, including availability, recovery objectives, policy conformance, speed to production for AI use cases, and trends in unit cost. This approach keeps resilience and value advancing in tandem.

Change should be managed with intent. Leaders should communicate what will be decommissioned and when, set clear expectations for new roles, and link reskilling efforts to visible wins such as reduced incident volume or faster time to production. Third-party risk and vendor resilience should be embedded into operations so that teams routinely test recovery and failover and treat dependencies as part of their core responsibilities rather than a peripheral concern.

Conclusion: Turning volatility into advantage

Resilience and value are not trade-offs. Hybrid-by-design enables enterprises to scale AI without breaking compliance, integrate acquisitions without chaos, and explain infrastructure choices in the language of EBITDA, margin, and time to market.

The path is practical. Run infrastructure with the business, not for it. Make placement decisions deliberate and transparent. Treat total cost as a unit-based input, not a back-end reconciliation. Segment estates for AI to secure capacity and compliance. Build durable capabilities in platform engineering, SRE, and AI operations. Embed governance, sustainability, and resilience into daily work.

Leaders who act rapidly by baselining TCO, standing up governance, and creating AI-ready zones are better positioned to establish the rhythm for durable modernization. They are poised to be able to scale AI responsibly, manage risk visibly, and deliver faster, cheaper, and more secure services than their peers. That is what it means to turn volatility into advantage.

Real estate

Real estate is now a first-order decision in data center strategy. More than just a shell, the building itself is an operating asset for an organization. Three groups are driving the market: private equity (PE), hyperscalers, and large enterprises. PE is moving from lease arbitrage to owning infrastructure. Hyperscalers continue to build for themselves and lock land, power, and fiber years ahead. A growing set of enterprises are building or co-owning facilities, often pairing them with onsite renewables or long-term power purchase agreements and using colocation for high-density AI. Chillers, air handlers, power distribution, and metering/submetering belong on the same scorecard as uptime, risk, and unit cost.

Site selection is about performance and risk. Start with demand and proximity to users and data to control latency. Filter by energy availability and price trajectory, grid interconnection timelines, and the ability to add low-carbon supply. Assess climate and water risk, permitting speed, labor depth, and regulatory geography. Plan for co-tenancy and future growth: what belongs on-premises, what fits in colocation, what sits at the edge. Operate the estate like a portfolio: common telemetry, a single pane of glass for leases and capacity, and service-level targets for both the plant and the compute it supports. Treat every site as part of one managed system so placement, contracting, and decommissioning decisions reinforce your AI strategy and the broader hybrid-by-design plan.

Appendix: Decision checklists and exhibits

Use these checklists and one-page exhibits to speed decisions and keep them auditable. They create a common view across technology, product, finance, and risk.

A. Top 25 workloads sheet (template fields)

- Service or product name, business owner, platform owner
- SLOs and recovery targets, criticality tier
- Data classification, residency and sovereignty notes
- Dependencies and blast radius map
- Current placement and rationale, target placement and rationale
- Current unit cost and trend, target unit cost and date
- Risk items, open exceptions to policy, mitigation plan
- Decommission or exit milestone, evidence required to close

B. Workload placement decision tree (criteria to step through)

Criterion	Choose…	If… (condition)	Notes/policy considerations
Data sensitivity and residency rules	On-premises or sovereign cloud region	Data is subject to strict regulatory, privacy, or residency controls (e.g., PHI, PII, or export restrictions).	Exception may be required if workload must integrate with external data pipelines.
	Public cloud region	Data is low sensitivity, and compliance allows cross-border or multi-tenant hosting.	Prefer compliant hyperscaler regions certified for required standards (e.g., HIPAA, FedRAMP).
Latency requirement	Edge deployment	Users, systems, or data sources require ultra-low latency (<10 ms).	Common for manufacturing, clinical instrumentation, or IoT scenarios.
	On-premises or colocation	Latency requirements are moderate (10–50 ms), and proximity to core systems is beneficial.	Often used for real-time control systems or enterprise resource planning integrated analytics.
	Cloud region	Latency tolerance is high (>50 ms), or workload is asynchronous.	Suitable for data analytics, machine learning (ML) training, and batch processing.
Availability and recovery objectives	Cloud region	High availability and global failover are priorities.	Leverage multi-region or multi-AZ capabilities.
	Colocation or on-premises	Recovery time objectives/recovery point objectives (RTOs/RPOs) are tight but under enterprise control.	Use redundant sites and enterprise disaster recovery tooling.

B. Workload placement decision tree (criteria to step through) (cont.)

Criterion	Choose…	If… (condition)	Notes/policy considerations
Traffic profile	Cloud region	Workload is seasonal or bursty, with variable compute needs.	Benefit from autoscaling and pay-per-use economics.
	On-premises or colocation	Traffic is steady-state and predictable.	Optimize for cost through fixed capacity investment.
AI pattern	Cloud region	Workload involves large-scale training or fine-tuning requiring elastic GPU/TPU resources.	Use managed ML platforms for efficiency and acceleration.
	On-premises	Workload is inference-heavy with stable, repeatable models serving internal users.	May reduce inference cost and maintain data control.
	Edge	Workload is embedded inference requiring real-time or disconnected operation.	Typically packaged in containers or device firmware.
Unit economics and budget ownership	Cloud region	Cost elasticity and variable consumption align with business unit budgets.	Monitor for sprawl; enforce tagging and chargeback.
	On-premises or colocation	Long-term utilization >70% and CapEx model is preferred by IT budget owner.	May justify hardware refresh cycle alignment.
Vendor resilience and exit requirements	Hybrid or multi-cloud	Exit risk, lock-in, or concentration concerns exist.	Abstract dependencies using containerization or orchestration layers.
	Single cloud region	Vendor risk is acceptable and standardization is prioritized.	Simplifies management and governance.

C. Source-of-service map (build, retrofit, or partner)

- Rows: GPU clusters, general compute, storage tiers, data platforms, networking and cross-connects, security and identity, observability, backup and disaster recovery
- Columns: build, retrofit, partner, with decision notes for control, security, spend transparency, speed to value, and exit path
- Add guardrails: telemetry access, portability of images and data, continuity testing cadence, SLA alignment to your SLOs

D. Unit economics before and after modernization (one-pager)

- Define unit: per transaction, per model call, per GB processed, or per user served
- Break down cost: compute, storage, network, platform services, labor operations, licenses
- Show current versus target with date and owner
- Note customer impact signals: latency, error rate, abandonment
- Link savings to a reinvestment item in the modernization backlog

E. AI zone reference exhibit

- Data adjacency plan and approved sources
- Model life cycle: training, fine-tuning, evaluation, promotion, rollback
- Security boundaries: network segmentation, identity, secrets, key management
- Observability stack: tracing, cost, performance, and policy conformance
- Capacity plan: accelerator profiles, power and cooling headroom, cross-connect design
- Recovery pattern and evidence required after each test cycle

How to use these

Keep all five artifacts in a shared location, review the workloads sheet and rate cards monthly, bring the placement tree and source-of-service map to the quarterly business review, and refresh the AI zone and unit economics exhibits during the annual reset. This keeps placement, cost, and risk decisions aligned and trackable as your footprint evolves.

Authors

Lou DiLorenzo
Tech, AI, & Data Strategy
US Practice Leader
Principal, Deloitte Consulting LLP
ldilorenzojr@deloitte.com

Jagjeet Gill
Principal, Deloitte Consulting LLP
jagjgill@deloitte.com

Heather Rangel
Principal, Deloitte Consulting LLP
hrangel@deloitte.com

Chris Thomas
US Hybrid Cloud Infrastructure Leader
Principal, Deloitte Consulting LLP
chrthomas@deloitte.com

Special thanks to Fay Chen Gerdes, Zack Grossenbacher, and Desmond Young for their contributions in writing this report.

Deloitte.

MAKING AN IMPACT THAT MATTERS since 1845